



US009626191B2

(12) **United States Patent**
Choquette et al.

(10) **Patent No.:** **US 9,626,191 B2**
(45) **Date of Patent:** **Apr. 18, 2017**

(54) **SHAPED REGISTER FILE READS**

(75) Inventors: **Jack Hilaire Choquette**, Palo Alto, CA (US); **Michael Fetterman**, Boxborough, MA (US); **Shirish Gadre**, Fremont, CA (US); **Xiaogang Qiu**, San Jose, CA (US); **Omkar Paranjape**, Austin, TX (US); **Anjana Rajendran**, San Jose, CA (US); **Stewart Glenn Carlton**, Madison, AL (US); **Eric Lyell Hill**, Palo Alto, CA (US); **Rajeshwaran Selvanesan**, Milpitas, CA (US); **Douglas J. Hahn**, Los Altos, CA (US)

(73) Assignee: **NVIDIA Corporation**, Santa Clara, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 1016 days.

(21) Appl. No.: **13/335,868**

(22) Filed: **Dec. 22, 2011**

(65) **Prior Publication Data**
US 2013/0166877 A1 Jun. 27, 2013

(51) **Int. Cl.**
G06F 9/38 (2006.01)
G06F 9/30 (2006.01)

(52) **U.S. Cl.**
CPC **G06F 9/3851** (2013.01); **G06F 9/3012** (2013.01)

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,658,551	B1 *	12/2003	Berenbaum et al.	712/24
7,339,592	B2 *	3/2008	Lindholm et al.	345/543
7,477,255	B1	1/2009	Lindholm et al.	
7,834,881	B2	11/2010	Liu et al.	
8,200,940	B1 *	6/2012	Lindholm	712/22
2009/0217020	A1	8/2009	Yourst	
2011/0078427	A1	3/2011	Shebanow et al.	
2011/0241744	A1 *	10/2011	Moudgill	G06F 9/30116 327/202

FOREIGN PATENT DOCUMENTS

TW	201019262	A	5/2010
TW	2010044311	A	12/2010

* cited by examiner

Primary Examiner — Eric Coleman

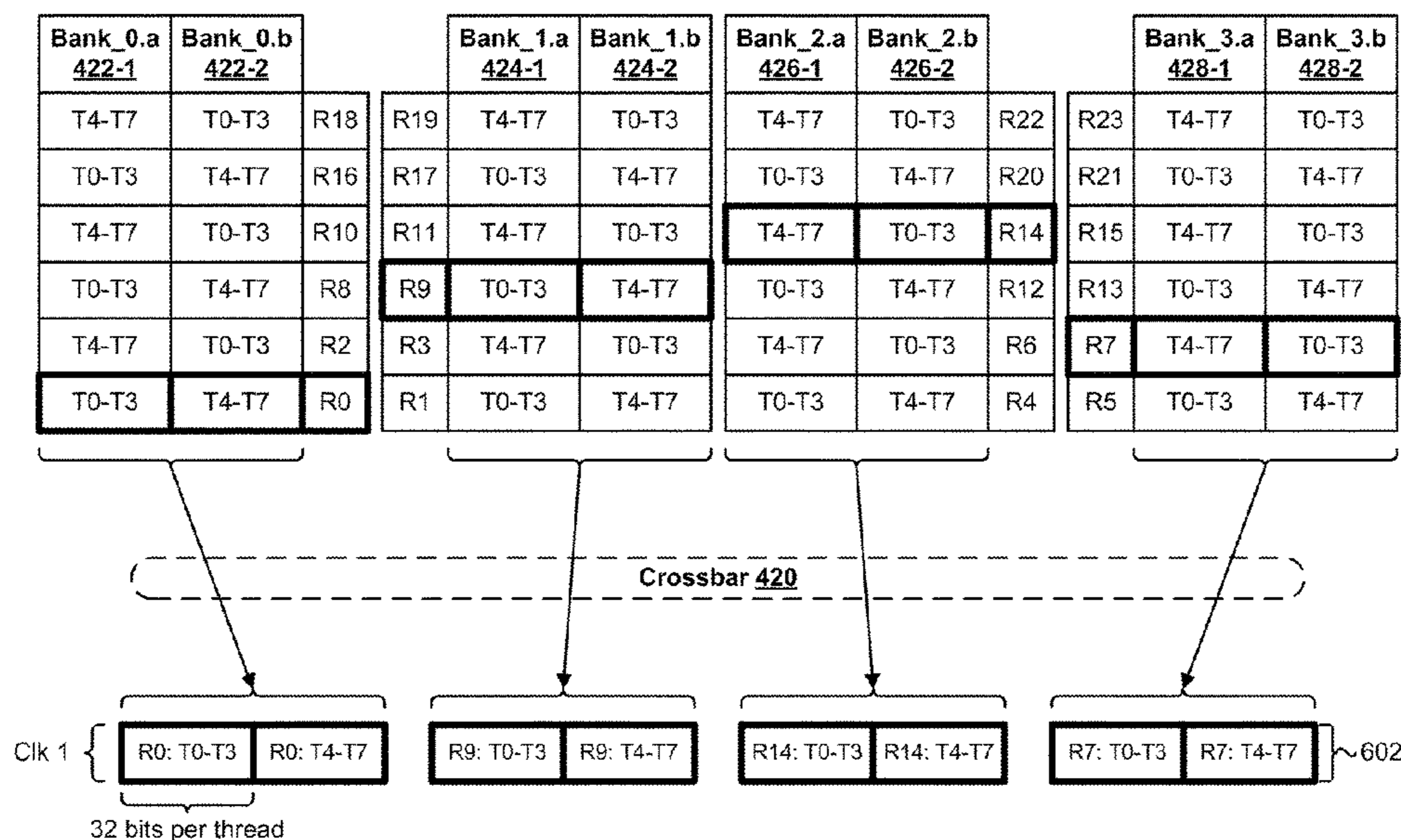
(74) *Attorney, Agent, or Firm* — Artega Law Group, LLP

(57) **ABSTRACT**

One embodiment of the present invention sets forth a technique for performing a shaped access of a register file that includes a set of N registers, wherein N is greater than or equal to two. The technique involves, for at least one thread included in a group of threads, receiving a request to access a first amount of data from each register in the set of N registers, and configuring a crossbar to allow the at least one thread to access the first amount of data from each register in the set of N registers.

20 Claims, 10 Drawing Sheets

32-bit reads of registers R0, R9, R14 and R7 for eight threads T0-T7



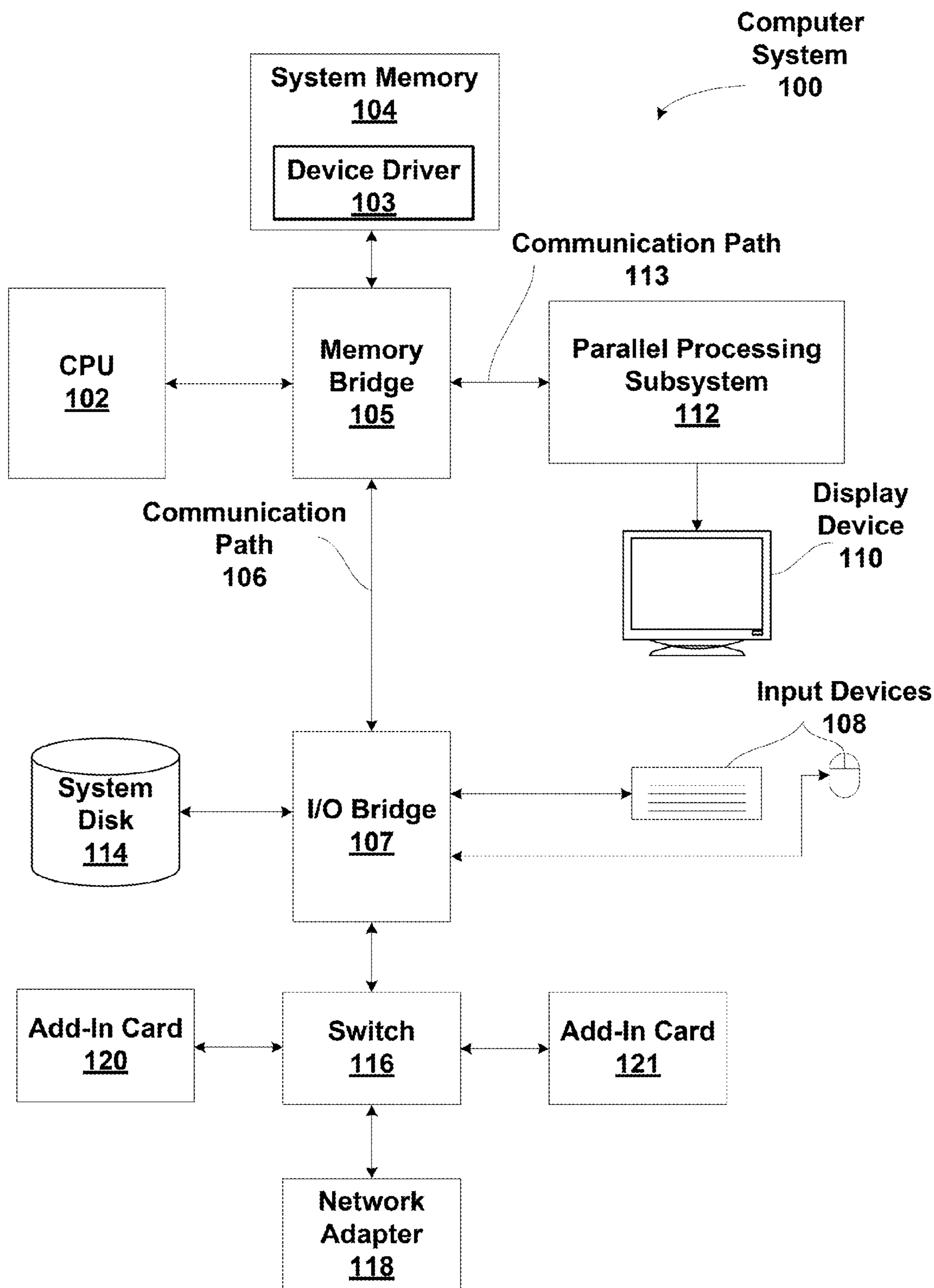


Figure 1

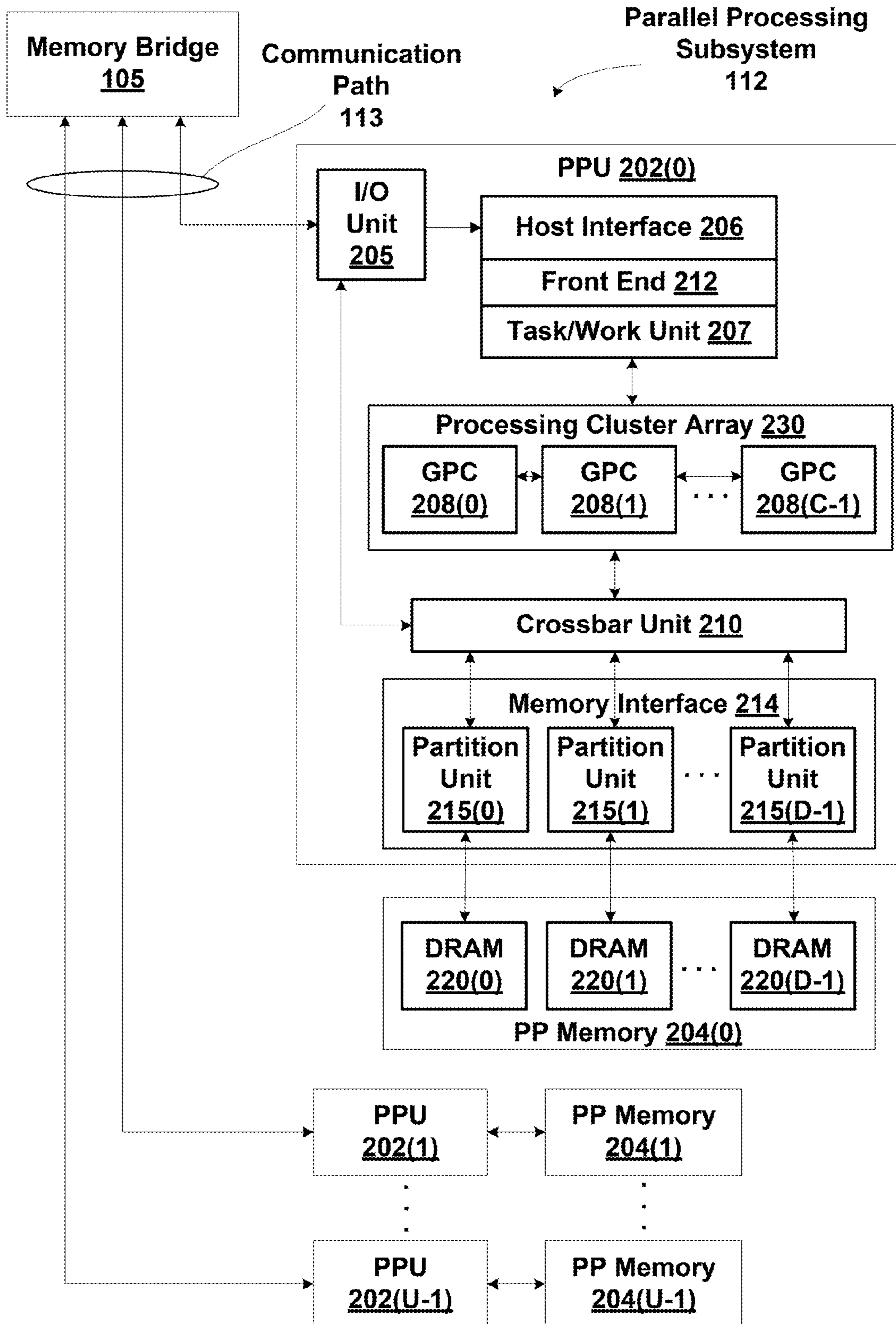


Figure 2

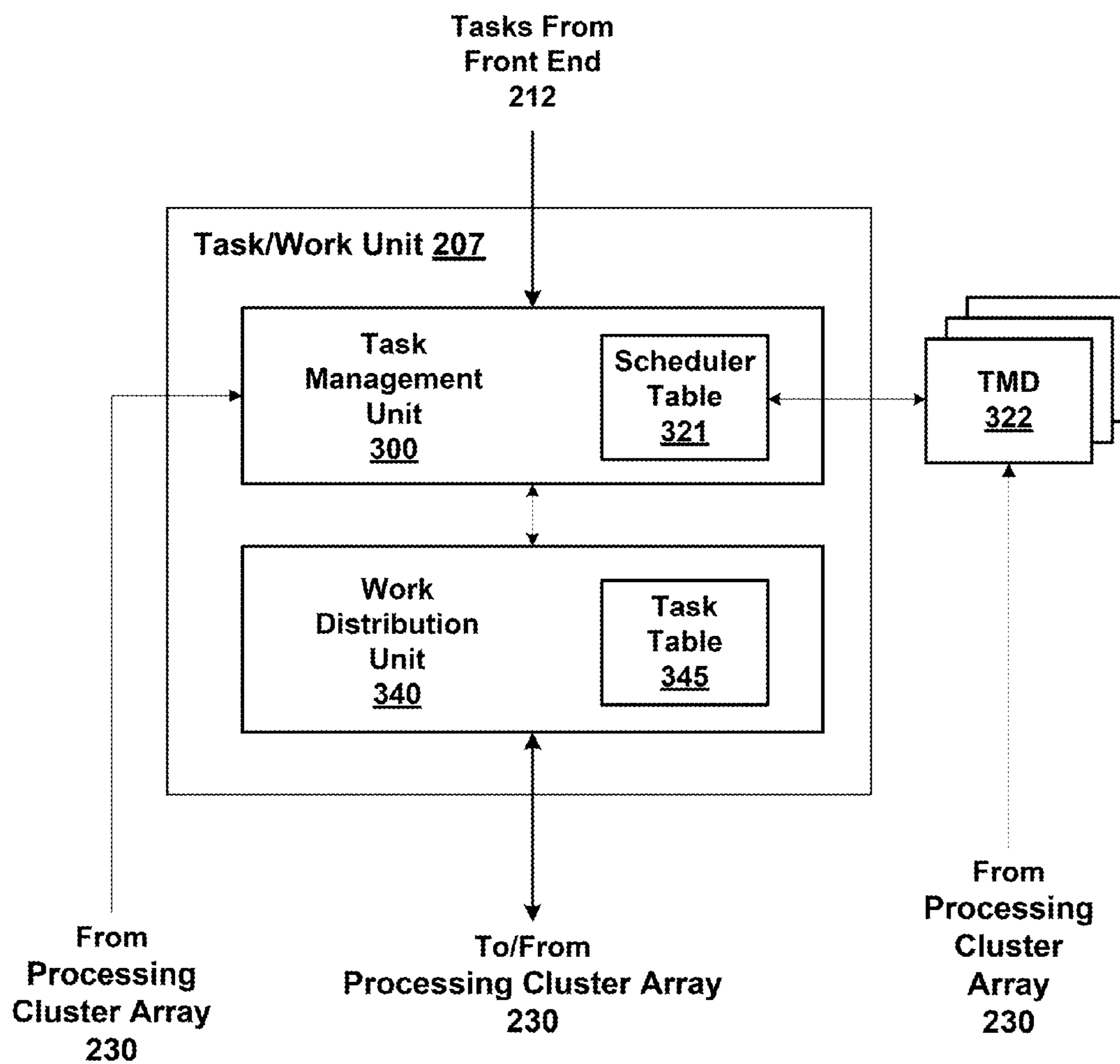


Figure 3A

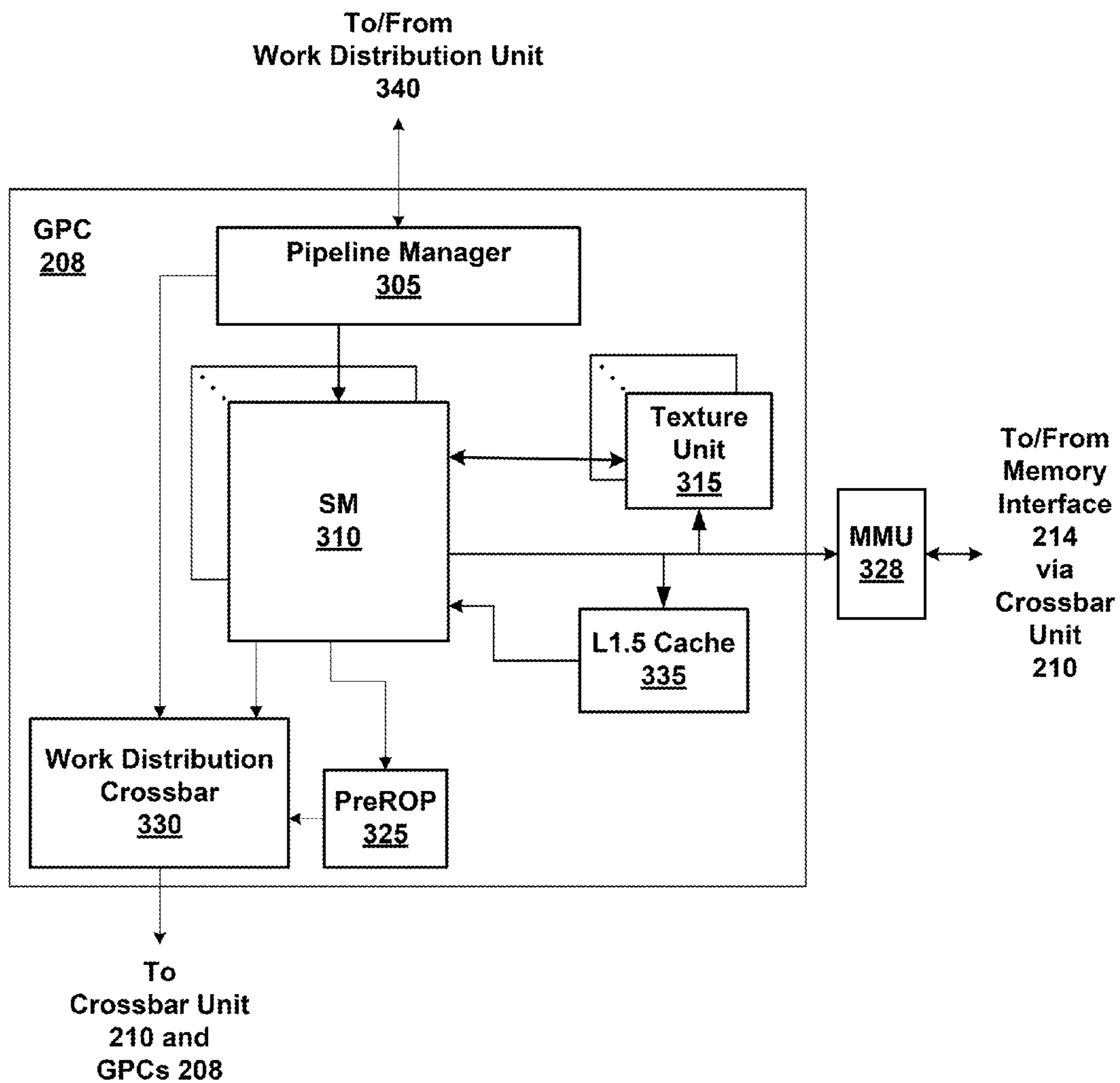


Figure 3B

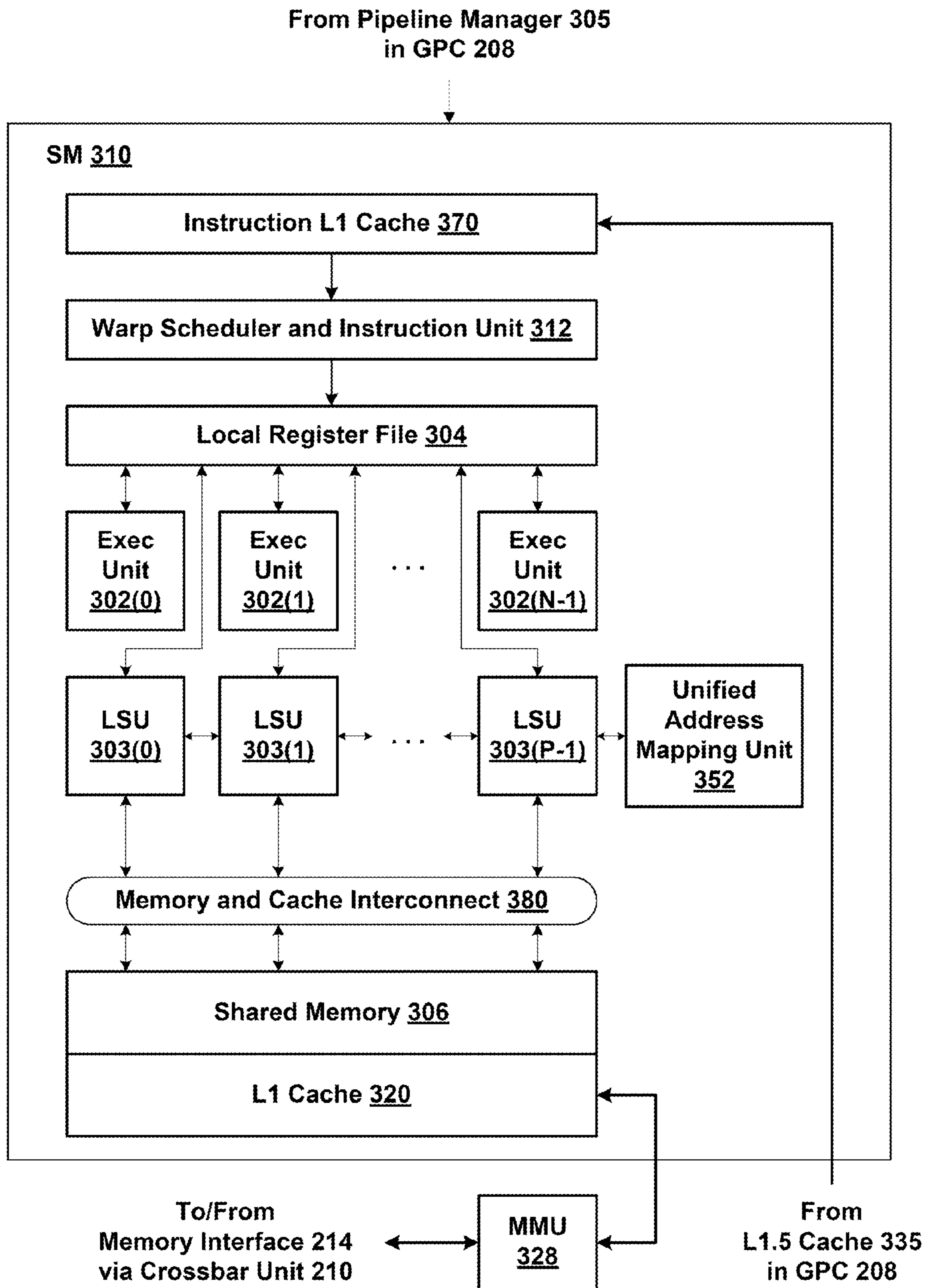


Figure 3C

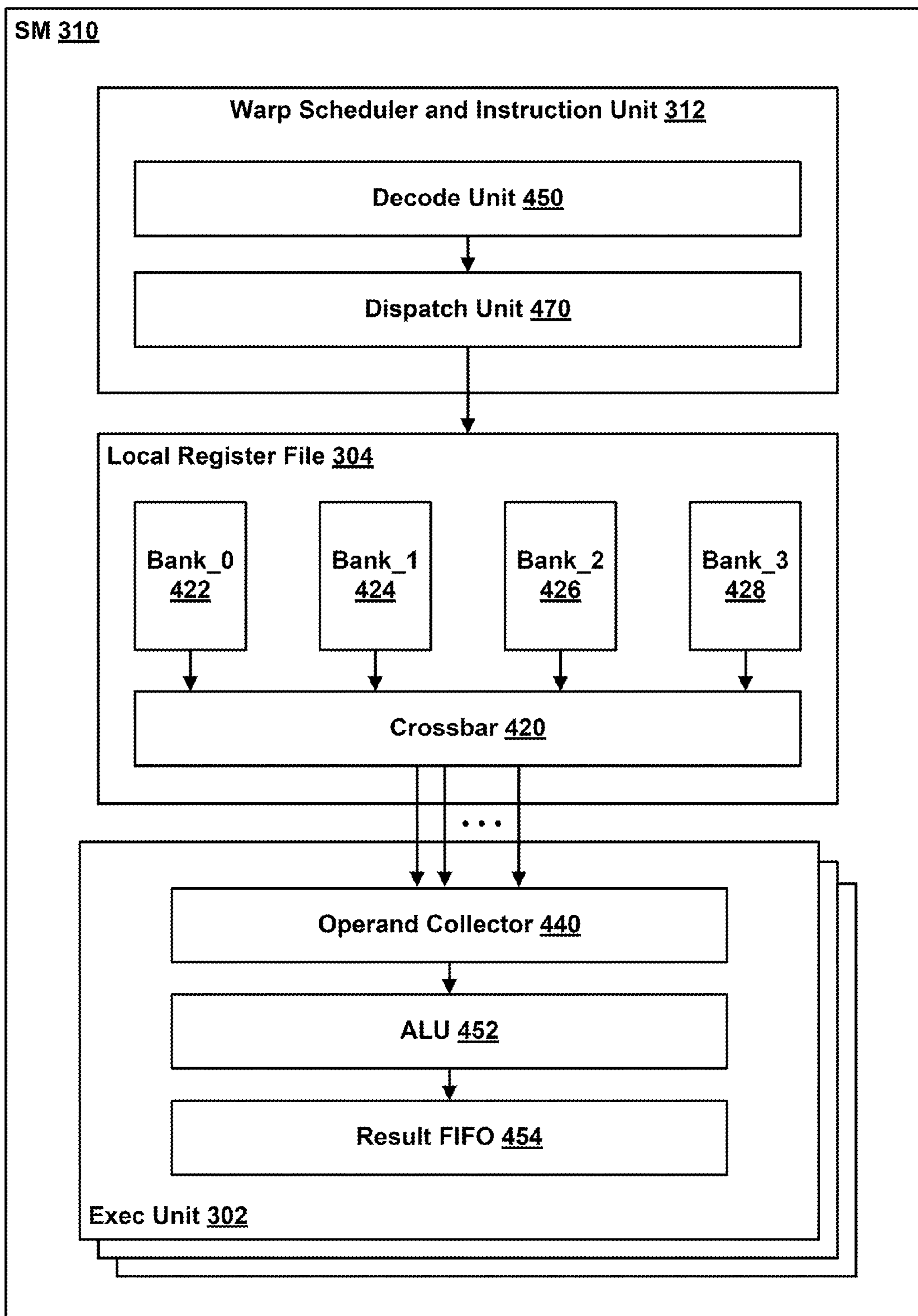


Figure 4

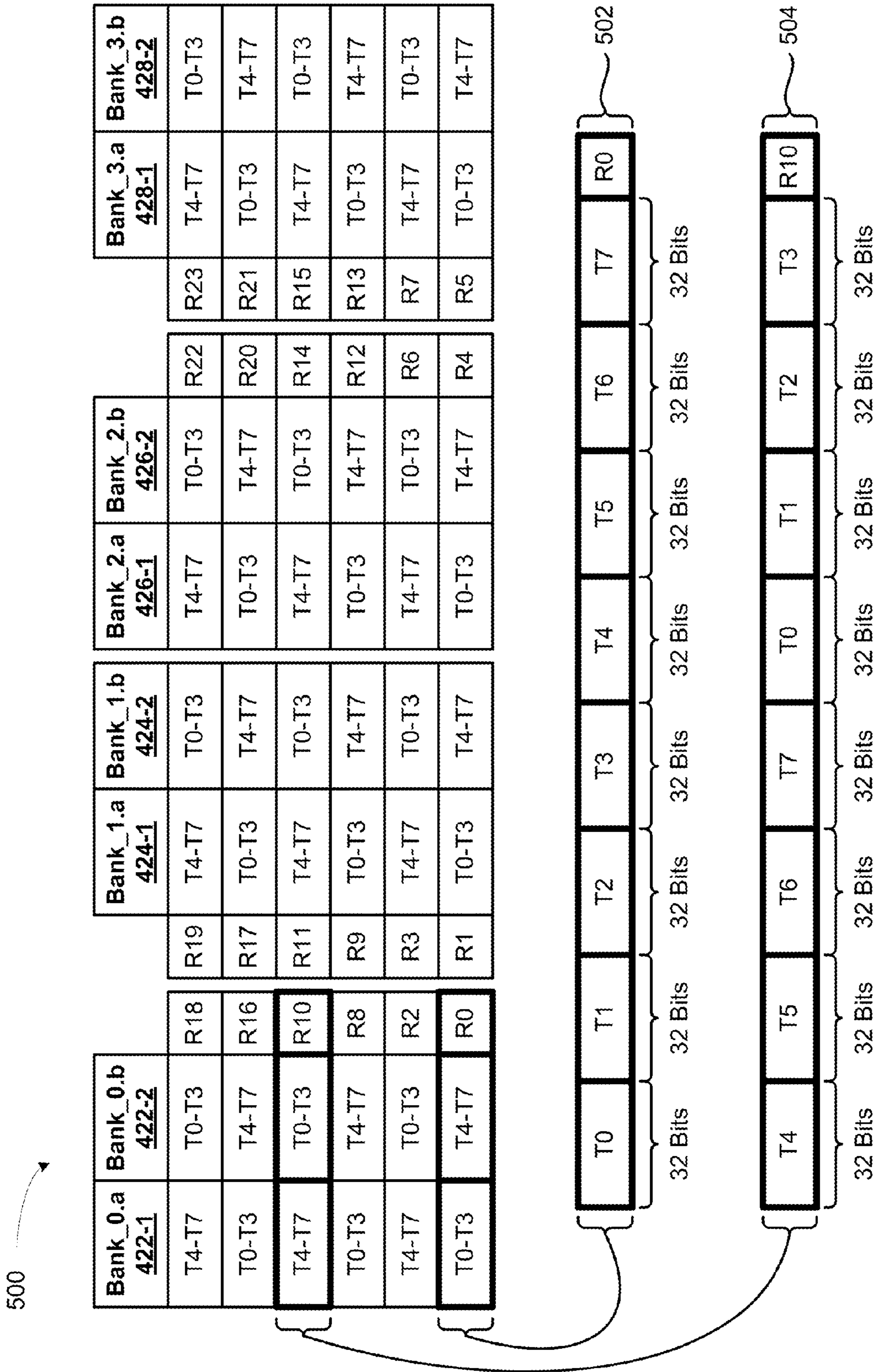


Figure 5

32-bit reads of registers R0, R9, R14 and R7 for eight threads T0-T7

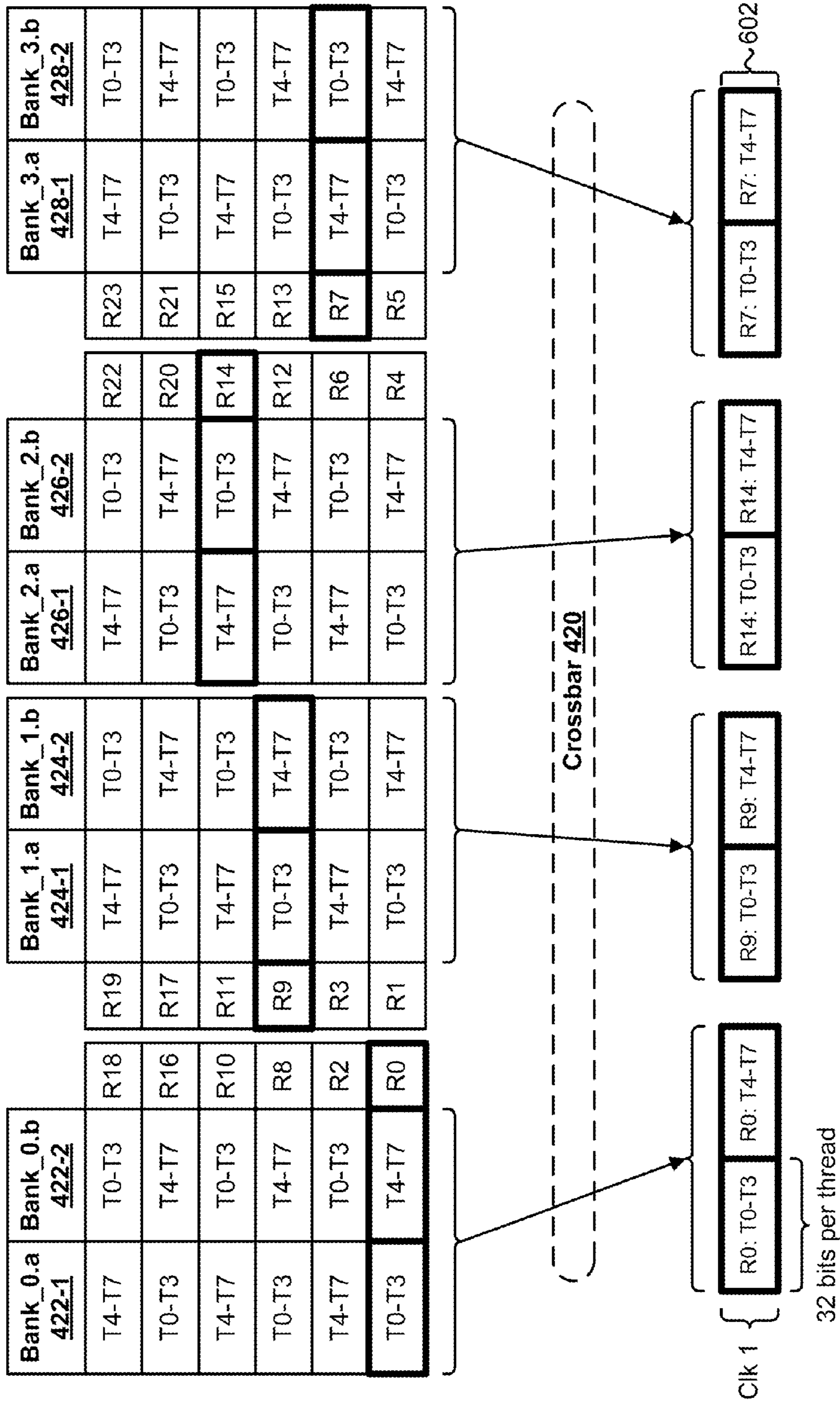


Figure 6



Figure 7

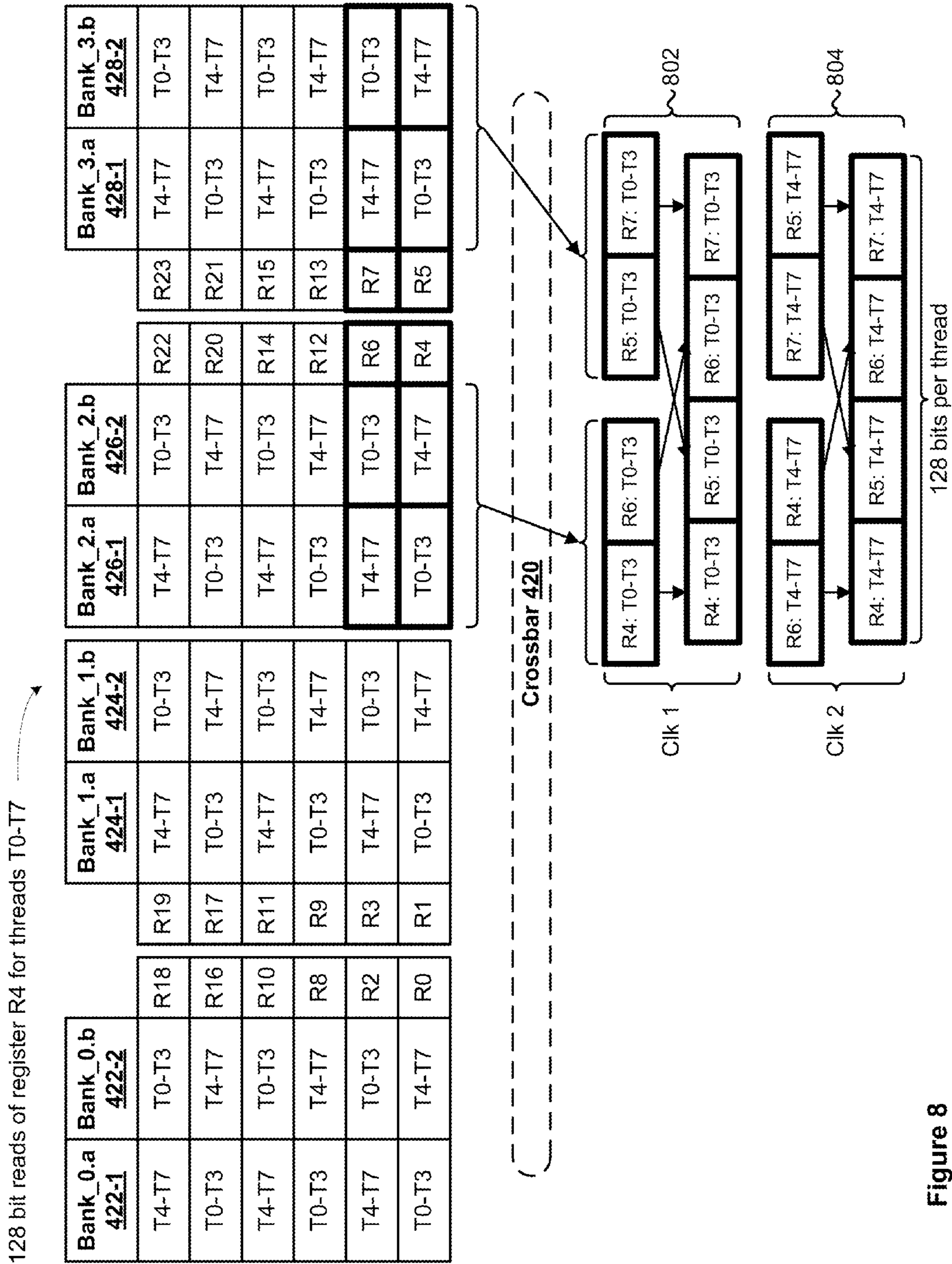


Figure 8

1

SHAPED REGISTER FILE READS

BACKGROUND OF THE INVENTION

Field of the Invention

The present disclosure generally relates to processor register files and, more specifically, to methods and apparatus for shaped register file reads.

Description of the Related Art

Parallel processors have multiple independent cores that enable multiple threads to be executed simultaneously using different hardware resources. SIMD (single instruction, multiple data) architecture processors execute the same instruction on each of the multiple cores where each core executes on different input data. MIMD (multiple instruction, multiple data) architecture processors execute different instructions on different cores with different input data supplied to each core. Parallel processors may also be multi-threaded, which enables two or more threads to execute substantially simultaneously using the resources of a single processing core (i.e., the different threads are executed on the core during different clock cycles).

When a processor schedules an instruction for execution by a processor core, the processor writes certain values into special registers in a register file coupled to the processor core. One register may store the opcode that specifies the operation to be performed by the processor core and additional registers may store operand values used as input to the processor core for executing the instruction. In order for an operation to be executed, each of the values must be written into the register file and then coupled to the inputs of the datapath via a crossbar or other data transmission means.

Oftentimes, an instruction for a thread refers to 32-bit, 64-bit or even 128-bit operands that are to be read from the register file. Conventional register files, however—which typically include a plurality of 32-bit slots—require the processor to transpose multiple 32-bit values read from the 32-bit slots into the 64-bit or 128-bit values requested by the thread, which requires several clock cycles to complete. One solution to this problem involves simply implementing register files that include larger slots, e.g., 64-bit slots. Unfortunately, such register files come at a much higher cost and increase the overall complexity of the processor in which they are included.

Accordingly, what is needed in the art is an improved technique for handling variable-sized data reads of a register file.

SUMMARY OF THE INVENTION

One embodiment of the present invention sets forth a method for performing a shaped access of a register file that includes a set of N registers, wherein N is greater than or equal to two. The method includes the steps of, for at least one thread included in a group of threads, receiving a request to access a first amount of data from each register in the set of N registers, and configuring a crossbar to allow the at least one thread to access the first amount of data from each register in the set of N registers.

BRIEF DESCRIPTION OF THE DRAWINGS

So that the manner in which the above recited features of the present disclosure can be understood in detail, a more particular description, briefly summarized above, may be had by reference to example embodiments, some of which are illustrated in the appended drawings. It is to be noted,

2

however, that the appended drawings illustrate only typical embodiments of this disclosure and are therefore not to be considered limiting of its scope, for the disclosure may admit to other equally effective embodiments.

FIG. 1 is a block diagram illustrating a computer system configured to implement one or more aspects of the present disclosure.

FIG. 2 is a block diagram of a parallel processing subsystem for the computer system of FIG. 1, according to one embodiment of the present disclosure.

FIG. 3A is a block diagram of the front end of FIG. 2, according to one embodiment of the present disclosure.

FIG. 3B is a block diagram of a general processing cluster within one of the parallel processing units of FIG. 2, according to one embodiment of the present disclosure.

FIG. 3C is a block diagram of a portion of the streaming multiprocessor of FIG. 3B, according to one embodiment of the present disclosure.

FIG. 4 is a block diagram of the streaming multiprocessor of FIG. 3B, according to another example embodiment of the present disclosure.

FIG. 5 illustrates a detailed view of the local register file of FIG. 3C, according to one embodiment of the present disclosure.

FIG. 6 illustrates 32-bit reads of four different registers for eight threads, according to one embodiment of the present disclosure.

FIG. 7 illustrates 64-bit reads of four different registers for eight threads, according to one embodiment of the present disclosure.

FIG. 8 illustrates 128-bit reads of a register for eight threads, according to one embodiment of the present disclosure.

DETAILED DESCRIPTION

In the following description, numerous specific details are set forth to provide a more thorough understanding of the present disclosure. However, it will be apparent to one of skill in the art that the present disclosure may be practiced without one or more of these specific details.

The disclosure describes methods and apparatus for source operand collector caching. In one embodiment, a processor includes a register file that may be coupled to storage elements (i.e., an operand collector) that provide inputs to the datapath of the processor core for executing an instruction. In order to reduce bandwidth between the register file and the operand collector, operands may be cached and reused in subsequent instructions. Consequently, only a subset of the operands specified by a given instruction may need to be loaded into the operand collector. A scheduling unit maintains a cache table for monitoring the register values currently stored in the operand collector. The scheduling unit may also configure the operand collector to select the particular storage elements that are coupled to the inputs to the datapath for a given instruction, allowing operands for two or more instructions to be cached concurrently.

System Overview

FIG. 1 is a block diagram illustrating a computer system 100 configured to implement one or more aspects of the present disclosure. Computer system 100 includes a central processing unit (CPU) 102 and a system memory 104 communicating via an interconnection path that may include a memory bridge 105. Memory bridge 105, which may be, e.g., a Northbridge chip, is connected via a bus or other

communication path **106** (e.g., a HyperTransport link) to an I/O (input/output) bridge **107**. I/O bridge **107**, which may be, e.g., a Southbridge chip, receives user input from one or more user input devices **108** (e.g., keyboard, mouse) and forwards the input to CPU **102** via communication path **106** and memory bridge **105**. A parallel processing subsystem **112** is coupled to memory bridge **105** via a bus or second communication path **113** (e.g., a Peripheral Component Interconnect (PCI) Express, Accelerated Graphics Port, or HyperTransport link); in one embodiment parallel processing subsystem **112** is a graphics subsystem that delivers pixels to a display device **110** (e.g., a conventional cathode ray tube or liquid crystal display based monitor). A system disk **114** is also connected to I/O bridge **107**. A switch **116** provides connections between I/O bridge **107** and other components such as a network adapter **118** and various add-in cards **120** and **121**. Other components (not explicitly shown), including universal serial bus (USB) or other port connections, compact disc (CD) drives, digital video disc (DVD) drives, film recording devices, and the like, may also be connected to I/O bridge **107**. The various communication paths shown in FIG. 1, including the specifically named communications paths **106** and **113**, may be implemented using any suitable protocols, such as PCI Express, AGP (Accelerated Graphics Port), HyperTransport, or any other bus or point-to-point communication protocol(s), and connections between different devices may use different protocols as is known in the art.

In one embodiment, the parallel processing subsystem **112** incorporates circuitry optimized for graphics and video processing, including, for example, video output circuitry, and constitutes a graphics processing unit (GPU). In another embodiment, the parallel processing subsystem **112** incorporates circuitry optimized for general purpose processing, while preserving the underlying computational architecture, described in greater detail herein. In yet another embodiment, the parallel processing subsystem **112** may be integrated with one or more other system elements in a single subsystem, such as joining the memory bridge **105**, CPU **102**, and I/O bridge **107** to form a system on chip (SoC).

It will be appreciated that the system shown herein is illustrative and that variations and modifications are possible. The connection topology, including the number and arrangement of bridges, the number of CPUs **102**, and the number of parallel processing subsystems **112**, may be modified as desired. For instance, in some embodiments, system memory **104** is connected to CPU **102** directly rather than through a bridge, and other devices communicate with system memory **104** via memory bridge **105** and CPU **102**. In other alternative topologies, parallel processing subsystem **112** is connected to I/O bridge **107** or directly to CPU **102**, rather than to memory bridge **105**. In still other embodiments, I/O bridge **107** and memory bridge **105** might be integrated into a single chip instead of existing as one or more discrete devices. Large embodiments may include two or more CPUs **102** and two or more parallel processing systems **112**. The particular components shown herein are optional; for instance, any number of add-in cards or peripheral devices might be supported. In some embodiments, switch **116** is eliminated, and network adapter **118** and add-in cards **120**, **121** connect directly to I/O bridge **107**.

FIG. 2 illustrates a parallel processing subsystem **112**, according to one embodiment of the present disclosure. As shown, parallel processing subsystem **112** includes one or more parallel processing units (PPUs) **202**, each of which is coupled to a local parallel processing (PP) memory **204**. In general, a parallel processing subsystem includes a number

U of PPUs, where $U \geq 1$. (Herein, multiple instances of like objects are denoted with reference numbers identifying the object and parenthetical numbers identifying the instance where needed.) PPUs **202** and parallel processing memories **204** may be implemented using one or more integrated circuit devices, such as programmable processors, application specific integrated circuits (ASICs), or memory devices, or in any other technically feasible fashion.

Referring again to FIG. 1 as well as FIG. 2, in some embodiments, some or all of PPUs **202** in parallel processing subsystem **112** are graphics processors with rendering pipelines that can be configured to perform various operations related to generating pixel data from graphics data supplied by CPU **102** and/or system memory **104** via memory bridge **105** and the second communication path **113**, interacting with local parallel processing memory **204** (which can be used as graphics memory including, e.g., a conventional frame buffer) to store and update pixel data, delivering pixel data to display device **110**, and the like. In some embodiments, parallel processing subsystem **112** may include one or more PPUs **202** that operate as graphics processors and one or more other PPUs **202** that are used for general-purpose computations. The PPUs may be identical or different, and each PPU may have a dedicated parallel processing memory device(s) or no dedicated parallel processing memory device(s). One or more PPUs **202** in parallel processing subsystem **112** may output data to display device **110** or each PPU **202** in parallel processing subsystem **112** may output data to one or more display devices **110**.

In operation, CPU **102** is the master processor of computer system **100**, controlling and coordinating operations of other system components. In particular, CPU **102** issues commands that control the operation of PPUs **202**. In some embodiments, CPU **102** writes a stream of commands for each PPU **202** to a data structure (not explicitly shown in either FIG. 1 or FIG. 2) that may be located in system memory **104**, parallel processing memory **204**, or another storage location accessible to both CPU **102** and PPU **202**. A pointer to each data structure is written to a pushbuffer to initiate processing of the stream of commands in the data structure. The PPU **202** reads command streams from one or more pushbuffers and then executes commands asynchronously relative to the operation of CPU **102**. Execution priorities may be specified for each pushbuffer by an application program via the device driver **103** to control scheduling of the different pushbuffers.

Referring back now to FIG. 2 as well as FIG. 1, each PPU **202** includes an I/O (input/output) unit **205** that communicates with the rest of computer system **100** via communication path **113**, which connects to memory bridge **105** (or, in one alternative embodiment, directly to CPU **102**). The connection of PPU **202** to the rest of computer system **100** may also be varied. In some embodiments, parallel processing subsystem **112** is implemented as an add-in card that can be inserted into an expansion slot of computer system **100**. In other embodiments, a PPU **202** can be integrated on a single chip with a bus bridge, such as memory bridge **105** or I/O bridge **107**. In still other embodiments, some or all elements of PPU **202** may be integrated on a single chip with CPU **102**.

In one embodiment, communication path **113** is a PCI Express link, in which dedicated lanes are allocated to each PPU **202**, as is known in the art. Other communication paths may also be used. An I/O unit **205** generates packets (or other signals) for transmission on communication path **113** and also receives all incoming packets (or other signals)

from communication path 113, directing the incoming packets to appropriate components of PPU 202. For example, commands related to processing tasks may be directed to a host interface 206, while commands related to memory operations (e.g., reading from or writing to parallel processing memory 204) may be directed to a memory crossbar unit 210. Host interface 206 reads each pushbuffer and outputs the command stream stored in the pushbuffer to a front end 212.

Each PPU 202 advantageously implements a highly parallel processing architecture. As shown in detail, PPU 202(0) includes a processing cluster array 230 that includes a number C of general processing clusters (GPCs) 208, where $C \geq 1$. Each GPC 208 is capable of executing a large number (e.g., hundreds or thousands) of threads concurrently, where each thread is an instance of a program. In various applications, different GPCs 208 may be allocated for processing different types of programs or for performing different types of computations. The allocation of GPCs 208 may vary dependent on the workload arising for each type of program or computation.

GPCs 208 receive processing tasks to be executed from a work distribution unit within a task/work unit 207. The work distribution unit receives pointers to processing tasks that are encoded as task metadata (TMD) and stored in memory. The pointers to TMDs are included in the command stream that is stored as a pushbuffer and received by the front end unit 212 from the host interface 206. Processing tasks that may be encoded as TMDs include indices of data to be processed, as well as state parameters and commands defining how the data is to be processed (e.g., what program is to be executed). The task/work unit 207 receives tasks from the front end 212 and ensures that GPCs 208 are configured to a valid state before the processing specified by each one of the TMDs is initiated. A priority may be specified for each TMD that is used to schedule execution of the processing task. Processing tasks can also be received from the processing cluster array 230. Optionally, the TMD can include a parameter that controls whether the TMD is added to the head or the tail for a list of processing tasks (or list of pointers to the processing tasks), thereby providing another level of control over priority.

Memory interface 214 includes a number D of partition units 215 that are each directly coupled to a portion of parallel processing memory 204, where $D \geq 1$. As shown, the number of partition units 215 generally equals the number of dynamic random access memory (DRAM) 220. In other embodiments, the number of partition units 215 may not equal the number of memory devices. Persons of ordinary skill in the art will appreciate that DRAM 220 may be replaced with other suitable storage devices and can be of generally conventional design. A detailed description is therefore omitted. Render targets, such as frame buffers or texture maps may be stored across DRAMs 220, allowing partition units 215 to write portions of each render target in parallel to efficiently use the available bandwidth of parallel processing memory 204.

Any one of GPCs 208 may process data to be written to any of the DRAMs 220 within parallel processing memory 204. Crossbar unit 210 is configured to route the output of each GPC 208 to the input of any partition unit 215 or to another GPC 208 for further processing. GPCs 208 communicate with memory interface 214 through crossbar unit 210 to read from or write to various external memory devices. In one embodiment, crossbar unit 210 has a connection to memory interface 214 to communicate with I/O unit 205, as well as a connection to local parallel processing

memory 204, thereby enabling the processing cores within the different GPCs 208 to communicate with system memory 104 or other memory that is not local to PPU 202. In the embodiment shown in FIG. 2, crossbar unit 210 is directly connected with I/O unit 205. Crossbar unit 210 may use virtual channels to separate traffic streams between the GPCs 208 and partition units 215.

Again, GPCs 208 can be programmed to execute processing tasks relating to a wide variety of applications, including but not limited to, linear and nonlinear data transforms, filtering of video and/or audio data, modeling operations (e.g., applying laws of physics to determine position, velocity and other attributes of objects), image rendering operations (e.g., tessellation shader, vertex shader, geometry shader, and/or pixel shader programs), and so on. PPUs 202 may transfer data from system memory 104 and/or local parallel processing memories 204 into internal (on-chip) memory, process the data, and write result data back to system memory 104 and/or local parallel processing memories 204, where such data can be accessed by other system components, including CPU 102 or another parallel processing subsystem 112.

A PPU 202 may be provided with any amount of local parallel processing memory 204, including no local memory, and may use local memory and system memory in any combination. For instance, a PPU 202 can be a graphics processor in a unified memory architecture (UMA) embodiment. In such embodiments, little or no dedicated graphics (parallel processing) memory would be provided, and PPU 202 would use system memory exclusively or almost exclusively. In UMA embodiments, a PPU 202 may be integrated into a bridge chip or processor chip or provided as a discrete chip with a high-speed link (e.g., PCI Express) connecting the PPU 202 to system memory via a bridge chip or other communication means.

As noted above, any number of PPUs 202 can be included in a parallel processing subsystem 112. For instance, multiple PPUs 202 can be provided on a single add-in card, or multiple add-in cards can be connected to communication path 113, or one or more of PPUs 202 can be integrated into a bridge chip. PPUs 202 in a multi-PPU system may be identical to or different from one another. For instance, different PPUs 202 might have different numbers of processing cores, different amounts of local parallel processing memory, and so on. Where multiple PPUs 202 are present, those PPUs may be operated in parallel to process data at a higher throughput than is possible with a single PPU 202. Systems incorporating one or more PPUs 202 may be implemented in a variety of configurations and form factors, including desktop, laptop, or handheld personal computers, servers, workstations, game consoles, embedded systems, and the like.

Multiple Concurrent Task Scheduling

Multiple processing tasks may be executed concurrently on the GPCs 208 and a processing task may generate one or more "child" processing tasks during execution. The task/work unit 207 receives the tasks and dynamically schedules the processing tasks and child processing tasks for execution by the GPCs 208.

FIG. 3A is a block diagram of the task/work unit 207 of FIG. 2, according to one embodiment of the present disclosure. The task/work unit 207 includes a task management unit 300 and the work distribution unit 340. The task management unit 300 organizes tasks to be scheduled based on execution priority levels. For each priority level, the task

management unit **300** stores a list of pointers to the TMDs **322** corresponding to the tasks in the scheduler table **321**, where the list may be implemented as a linked list. The TMDs **322** may be stored in the PP memory **204** or system memory **104**. The rate at which the task management unit **300** accepts tasks and stores the tasks in the scheduler table **321** is decoupled from the rate at which the task management unit **300** schedules tasks for execution. Therefore, the task management unit **300** may collect several tasks before scheduling the tasks. The collected tasks may then be scheduled based on priority information or using other techniques, such as round-robin scheduling.

The work distribution unit **340** includes a task table **345** with slots that may each be occupied by the TMD **322** for a task that is being executed. The task management unit **300** may schedule tasks for execution when there is a free slot in the task table **345**. When there is not a free slot, a higher priority task that does not occupy a slot may evict a lower priority task that does occupy a slot. When a task is evicted, the task is stopped, and if execution of the task is not complete, then a pointer to the task is added to a list of task pointers to be scheduled so that execution of the task will resume at a later time. When a child processing task is generated, during execution of a task, a pointer to the child task is added to the list of task pointers to be scheduled. A child task may be generated by a TMD **322** executing in the processing cluster array **230**.

Unlike a task that is received by the task/work unit **207** from the front end **212**, child tasks are received from the processing cluster array **230**. Child tasks are not inserted into pushbuffers or transmitted to the front end. The CPU **102** is not notified when a child task is generated or data for the child task is stored in memory. Another difference between the tasks that are provided through pushbuffers and child tasks is that the tasks provided through the pushbuffers are defined by the application program whereas the child tasks are dynamically generated during execution of the tasks.

Task Processing Overview

FIG. 3B is a block diagram of a GPC **208** within one of the PPUs **202** of FIG. 2, according to one embodiment of the present disclosure. Each GPC **208** may be configured to execute a large number of threads in parallel, where the term “thread” refers to an instance of a particular program executing on a particular set of input data. In some embodiments, single-instruction, multiple-data (SIMD) instruction issue techniques are used to support parallel execution of a large number of threads without providing multiple independent instruction units. In other embodiments, single-instruction, multiple-thread (SIMT) techniques are used to support parallel execution of a large number of generally synchronized threads, using a common instruction unit configured to issue instructions to a set of processing engines within each one of the GPCs **208**. Unlike a SIMD execution regime, where all processing engines typically execute identical instructions, SIMT execution allows different threads to more readily follow divergent execution paths through a given thread program. Persons of ordinary skill in the art will understand that a SIMD processing regime represents a functional subset of a SIMT processing regime.

Operation of GPC **208** is advantageously controlled via a pipeline manager **305** that distributes processing tasks to streaming multiprocessors (SMs) **310**. Pipeline manager **305** may also be configured to control a work distribution crossbar **330** by specifying destinations for processed data output by SMs **310**.

In one embodiment, each GPC **208** includes a number M of SMs **310**, where $M \geq 1$, each SM **310** configured to process one or more thread groups. Also, each SM **310** advantageously includes an identical set of functional execution units (e.g., execution units and load-store units—shown as Exec units **302** and LSUs **303** in FIG. 3C) that may be pipelined, allowing a new instruction to be issued before a previous instruction has finished, as is known in the art. Any combination of functional execution units may be provided. In one embodiment, the functional units support a variety of operations including integer and floating point arithmetic (e.g., addition and multiplication), comparison operations, Boolean operations (AND, OR, XOR), bit-shifting, and computation of various algebraic functions (e.g., planar interpolation, trigonometric, exponential, and logarithmic functions, etc.); and the same functional unit hardware can be leveraged to perform different operations.

The series of instructions transmitted to a particular GPC **208** constitutes a thread, as previously defined herein, and the collection of a certain number of concurrently executing threads across the parallel processing engines (not shown) within an SM **310** is referred to herein as a “warp” or “thread group.” As used herein, a “thread group” refers to a group of threads concurrently executing the same program on different input data, with one thread of the group being assigned to a different processing engine within an SM **310**. A thread group may include fewer threads than the number of processing engines within the SM **310**, in which case some processing engines will be idle during cycles when that thread group is being processed. A thread group may also include more threads than the number of processing engines within the SM **310**, in which case processing will take place over consecutive clock cycles. Since each SM **310** can support up to G thread groups concurrently, it follows that up to $G \cdot M$ thread groups can be executing in GPC **208** at any given time.

Additionally, a plurality of related thread groups may be active (in different phases of execution) at the same time within an SM **310**. This collection of thread groups is referred to herein as a “cooperative thread array” (“CTA”) or “thread array.” The size of a particular CTA is equal to $m \cdot k$, where k is the number of concurrently executing threads in a thread group and is typically an integer multiple of the number of parallel processing engines within the SM **310**, and m is the number of thread groups simultaneously active within the SM **310**. The size of a CTA is generally determined by the programmer and the amount of hardware resources, such as memory or registers, available to the CTA.

Each SM **310** contains a level one (L1) cache (shown in FIG. 3C) or uses space in a corresponding L1 cache outside of the SM **310** that is used to perform load and store operations. Each SM **310** also has access to level two (L2) caches that are shared among all GPCs **208** and may be used to transfer data between threads. Finally, SMs **310** also have access to off-chip “global” memory, which can include, e.g., parallel processing memory **204** and/or system memory **104**. It is to be understood that any memory external to PPU **202** may be used as global memory. Additionally, a level one-point-five (L1.5) cache **335** may be included within the GPC **208**, configured to receive and hold data fetched from memory via memory interface **214** requested by SM **310**, including instructions, uniform data, and constant data, and provide the requested data to SM **310**. Embodiments having multiple SMs **310** in GPC **208** beneficially share common instructions and data cached in L1.5 cache **335**.

Each GPC **208** may include a memory management unit (MMU) **328** that is configured to map virtual addresses into physical addresses. In other embodiments, MMU(s) **328** may reside within the memory interface **214**. The MMU **328** includes a set of page table entries (PTEs) used to map a virtual address to a physical address of a tile and optionally a cache line index. The MMU **328** may include address translation lookaside buffers (TLB) or caches which may reside within multiprocessor SM **310** or the L1 cache or GPC **208**. The physical address is processed to distribute surface data access locality to allow efficient request interleaving among partition units **215**. The cache line index may be used to determine whether or not a request for a cache line is a hit or miss.

In graphics and computing applications, a GPC **208** may be configured such that each SM **310** is coupled to a texture unit **315** for performing texture mapping operations, e.g., determining texture sample positions, reading texture data, and filtering the texture data. Texture data is read from an internal texture L1 cache (not shown) or in some embodiments from the L1 cache within SM **310** and is fetched from an L2 cache that is shared between all GPCs **208**, parallel processing memory **204**, or system memory **104**, as needed. Each SM **310** outputs processed tasks to work distribution crossbar **330** in order to provide the processed task to another GPC **208** for further processing or to store the processed task in an L2 cache, parallel processing memory **204**, or system memory **104** via crossbar unit **210**. A preROP (pre-raster operations) **325** is configured to receive data from SM **310**, direct data to ROP units within partition units **215**, and perform optimizations for color blending, organize pixel color data, and perform address translations.

It will be appreciated that the core architecture described herein is illustrative and that variations and modifications are possible. Any number of processing units, e.g., SMs **310** or texture units **315**, preROPs **325** may be included within a GPC **208**. Further, as shown in FIG. 2, a PPU **202** may include any number of GPCs **208** that are advantageously functionally similar to one another so that execution behavior does not depend on which GPC **208** receives a particular processing task. Further, each GPC **208** advantageously operates independently of other GPCs **208** using separate and distinct processing units, L1 caches to execute tasks for one or more application programs.

Persons of ordinary skill in the art will understand that the architecture described in FIGS. 1, 2, 3A, and 3B in no way limits the scope of the present invention and that the techniques taught herein may be implemented on any properly configured processing unit, including, without limitation, one or more CPUs, one or more multi-core CPUs, one or more PPUs **202**, one or more GPCs **208**, one or more graphics or special purpose processing units, or the like, without departing the scope of the present invention.

In embodiments of the present invention, it is desirable to use PPU **202** or other processor(s) of a computing system to execute general-purpose computations using thread arrays. Each thread in the thread array is assigned a unique thread identifier (“thread ID”) that is accessible to the thread during the thread’s execution. The thread ID, which can be defined as a one-dimensional or multi-dimensional numerical value controls various aspects of the thread’s processing behavior. For instance, a thread ID may be used to determine which portion of the input data set a thread is to process and/or to determine which portion of an output data set a thread is to produce or write.

A sequence of per-thread instructions may include at least one instruction that defines a cooperative behavior between

the representative thread and one or more other threads of the thread array. For example, the sequence of per-thread instructions might include an instruction to suspend execution of operations for the representative thread at a particular point in the sequence until such time as one or more of the other threads reach that particular point, an instruction for the representative thread to store data in a shared memory to which one or more of the other threads have access, an instruction for the representative thread to atomically read and update data stored in a shared memory to which one or more of the other threads have access based on their thread IDs, or the like. The CTA program can also include an instruction to compute an address in the shared memory from which data is to be read, with the address being a function of a thread ID. By defining suitable functions and providing synchronization techniques, data can be written to a given location in shared memory by one thread of a CTA and read from that location by a different thread of the same CTA in a predictable manner. Consequently, any desired pattern of data sharing among threads can be supported, and any thread in a CTA can share data with any other thread in the same CTA. The extent, if any, of data sharing among threads of a CTA is determined by the CTA program; thus, it is to be understood that in a particular application that uses CTAs, the threads of a CTA might or might not actually share data with each other, depending on the CTA program, and the terms “CTA” and “thread array” are used synonymously herein.

FIG. 3C is a block diagram of the SM **310** of FIG. 3B, according to one embodiment of the present disclosure. The SM **310** includes an instruction L1 cache **370** that is configured to receive instructions and constants from memory via L1.5 cache **335**. A warp scheduler and instruction unit **312** receives instructions and constants from the instruction L1 cache **370** and controls local register file **304** and SM **310** functional units according to the instructions and constants. The SM **310** functional units include N exec (execution or processing) units **302** and P load-store units (LSU) **303**.

SM **310** provides on-chip (internal) data storage with different levels of accessibility. Special registers (not shown) are readable but not writeable by LSU **303** and are used to store parameters defining each thread’s “position.” In one embodiment, special registers include one register per thread (or per exec unit **302** within SM **310**) that stores a thread ID; each thread ID register is accessible only by a respective one of the exec unit **302**. Special registers may also include additional registers, readable by all threads that execute the same processing task represented by a TMD **322** (or by all LSUs **303**) that store a CTA identifier, the CTA dimensions, the dimensions of a grid to which the CTA belongs (or queue position if the TMD **322** encodes a queue task instead of a grid task), and an identifier of the TMD **322** to which the CTA is assigned.

If the TMD **322** is a grid TMD, execution of the TMD **322** causes a fixed number of CTAs to be launched and executed to process the fixed amount of data stored in the queue **525**. The number of CTAs is specified as the product of the grid width, height, and depth. The fixed amount of data may be stored in the TMD **322** or the TMD **322** may store a pointer to the data that will be processed by the CTAs. The TMD **322** also stores a starting address of the program that is executed by the CTAs.

If the TMD **322** is a queue TMD, then a queue feature of the TMD **322** is used, meaning that the amount of data to be processed is not necessarily fixed. Queue entries store data for processing by the CTAs assigned to the TMD **322**. The queue entries may also represent a child task that is gener-

ated by another TMD 322 during execution of a thread, thereby providing nested parallelism. Typically, execution of the thread, or CTA that includes the thread, is suspended until execution of the child task completes. The queue may be stored in the TMD 322 or separately from the TMD 322, in which case the TMD 322 stores a queue pointer to the queue. Advantageously, data generated by the child task may be written to the queue while the TMD 322 representing the child task is executing. The queue may be implemented as a circular queue so that the total amount of data is not limited to the size of the queue.

CTAs that belong to a grid have implicit grid width, height, and depth parameters indicating the position of the respective CTA within the grid. Special registers are written during initialization in response to commands received via front end 212 from device driver 103 and do not change during execution of a processing task. The front end 212 schedules each processing task for execution. Each CTA is associated with a specific TMD 322 for concurrent execution of one or more tasks. Additionally, a single GPC 208 may execute multiple tasks concurrently.

A parameter memory (not shown) stores runtime parameters (constants) that can be read but not written by any thread within the same CTA (or any LSU 303). In one embodiment, device driver 103 provides parameters to the parameter memory before directing SM 310 to begin execution of a task that uses these parameters. Any thread within any CTA (or any exec unit 302 within SM 310) can access global memory through a memory interface 214. Portions of global memory may be stored in the L1 cache 320.

Local register file 304 is used by each thread as scratch space; each register is allocated for the exclusive use of one thread, and data in any of local register file 304 is accessible only to the thread to which the register is allocated. Local register file 304 can be implemented as a register file that is physically or logically divided into P lanes, each having some number of entries (where each entry might store, e.g., a 32-bit word), which is described in detail below in conjunction with FIGS. 5-8. One lane is assigned to each of the N exec units 302 and P load-store units LSU 303, and corresponding entries in different lanes can be populated with data for different threads executing the same program to facilitate SIMD execution. Different portions of the lanes can be allocated to different ones of the G concurrent thread groups, so that a given entry in the local register file 304 is accessible only to a particular thread. In one embodiment, certain entries within the local register file 304 are reserved for storing thread identifiers, implementing one of the special registers. Additionally, a uniform L1 cache 375 stores uniform or constant values for each lane of the N exec units 302 and P load-store units LSU 303.

Shared memory 306 is accessible to threads within a single CTA; in other words, any location in shared memory 306 is accessible to any thread within the same CTA (or to any processing engine within SM 310). Shared memory 306 can be implemented as a shared register file or shared on-chip cache memory with an interconnect that allows any processing engine to read from or write to any location in the shared memory. In other embodiments, shared state space might map onto a per-CTA region of off-chip memory, and be cached in L1 cache 320. The parameter memory can be implemented as a designated section within the same shared register file or shared cache memory that implements shared memory 306, or as a separate shared register file or on-chip cache memory to which the LSUs 303 have read-only access. In one embodiment, the area that implements the parameter memory is also used to store the CTA ID and task

ID, as well as CTA and grid dimensions or queue position, implementing portions of the special registers. Each LSU 303 in SM 310 is coupled to a unified address mapping unit 352 that converts an address provided for load and store instructions that are specified in a unified memory space into an address in each distinct memory space. Consequently, an instruction may be used to access any of the local, shared, or global memory spaces by specifying an address in the unified memory space.

The L1 cache 320 in each SM 310 can be used to cache private per-thread local data and also per-application global data. In some embodiments, the per-CTA shared data may be cached in the L1 cache 320. The LSUs 303 are coupled to the shared memory 306 and the L1 cache 320 via a memory and cache interconnect 380.

Shaped Register File Reads

FIG. 4 is a block diagram of the SM 310 of FIG. 3B, according to another example embodiment of the present disclosure. Although not shown explicitly, SM 310 of FIG. 4 may contain some or all of the components of SM 310 of FIG. 3C, described above, in addition to the components explicitly shown in FIG. 4. As shown in FIG. 4, SM 310 includes the warp scheduler and instruction unit 312, the local register file 304, and one or more execution units 302. Warp scheduler and instruction unit 312 includes a decode unit 450 and a dispatch unit 470. Decode unit 450 receives a next instruction to be dispatched to execution unit 302. The decode unit 450 performs a full decode of the instruction and transmits the decoded instruction to the dispatch unit 470. For example, decode unit 450 will determine the particular type of instruction specified by an opcode in the instruction and the particular local register files 304 indices that are specified as operands to the instruction as well as a register index for storing the result of an instruction. In some embodiments, instructions may be dual or quad issued and decode unit 450 may implement separate decode logic for each issued instruction. Dispatch unit 470 implements a FIFO and writes the decoded values to local register file 304 for execution. In embodiments that issue multiple instructions simultaneously, dispatch unit 470 may issue each instruction to a different portion of the execution units 302 of SM 310.

In one embodiment, the local register file 304 includes four banks of registers (bank_0 422, bank_1 424, bank_2 426, and bank_3 428). In most conventional processing units, the local register file 304 may be quite small. For example, the x86 CPU architecture includes 8 (32-bit) or 16 (64-bit) registers per processor core. In contrast, each bank of local register file 304 may include a large number of registers available to use as inputs to the execution units 302, as described in detail below in conjunction with FIGS. 5-8. Crossbar 420 is configured to connect the various registers in each of the register banks to execution unit 302, which implements a datapath for performing an operation. In one embodiment, the datapath includes an operand collector 440, an arithmetic logic unit (ALU) 452 and a result FIFO 454. In one embodiment, operand collector 440 includes plurality of storage elements that may be coupled to the inputs of the ALU 452. Each of the storage elements may be a flip-flop, latch, or any other technically feasible circuit component capable of temporarily storing a value to supply as an input to the ALU 452. The outputs of the storage elements may be hardwired to the circuit elements that comprise the ALU 452 such as an adder circuit or a floating point multiplier circuit. As shown, the output from ALU 452 is input into the result

FIFO 454 such that arithmetic results may be reused by ALU 452 or subsequently stored back into memory, e.g., shared memory 306 via memory and cache interconnect 380.

In conventional processor designs that include relatively small register files, a crossbar or other interconnect may be configured to couple any register in a local register file to any one of the datapath inputs during a single clock cycle. Furthermore, in conventional processors, a number of registers, the number equal to the number of inputs to the datapath, may simultaneously be connected to any of the datapath inputs in a single clock cycle. It will be appreciated that increasing the size of the local register file 304 beyond 8 or 16 registers may increase the complexity of crossbar 420.

FIG. 5 illustrates a detailed view 500 of the local register file 403, according to one embodiment of the present disclosure. As shown in FIG. 5, the bank_0 (422) is separated into bank_0.a (422-1) and bank_0.b (422-2) sub-banks, the bank_1 (424) is separated into bank_1.a (424-1) and bank_1.b (424-2) sub-banks, the bank_2 (426) is separated into bank_2.a (426-1) and bank_2.b (426-2) sub-banks, and the bank_3 (428) is separated into bank_3.a (428-1) and bank_3.b (428-2) sub-banks. As also shown, each sub-bank includes six individually-addressable registers for eight different threads T0, T1, T2, T3, T4, T5, T6 and T7. For example, a separate and distinct 32-bit register R0 is included across bank_0.a (422-1) and bank_0.b (422-2) for each of the threads T0-T3 and T4-T7, respectively, which is illustrated by the register R0 breakdown 502. Of course, embodiments of the invention encompass each sub-bank being configured to store any number N of individually-addressable registers.

The location of registers for threads T0-T3 and threads T4-T7 is alternated across each evenly-numbered register that follows R0, e.g., a separate and distinct 32-bit register R10 is included across bank_0.a (422-1) and bank_0.b (422-2) for each of the threads T4-T7 and T0-T3, which is illustrated by register breakdown 504. Similarly, the location of registers for threads T0-T3 and T4-T7 is alternated across each odd-numbered register that follows R1. As is described in further detail below in conjunction with FIGS. 6-8, the banks 422, 424, 426 and 428 within the local register file 304 are organized according to the detailed view 500 in order to enable the threads to issue 32-bit, 64-bit and 128-bit reads of registers using only a single clock cycle.

It will be apparent to those having ordinary skill in the art that each register included in local register file 403 may be configured to store any number of bits, e.g., 8 bits, 16 bits, 32 bits, 64 bits, 128 bits, and so on. Moreover, the register assignments to sub-banks can be generalized to $N*2$ sub-banks when reading expanded values (e.g., the 64-bit reads described below in conjunction with FIG. 7), and can further be generalized to $N*4$ sub-banks when reading further-expanded values (e.g., the 128-bit reads described below in conjunction with FIG. 8), for arbitrary values of N.

FIG. 6 illustrates 32-bit reads of registers R0, R9, R14 and R7 for eight threads T0-T7, according to one embodiment of the present disclosure. In the example illustrated in FIG. 6, each of the threads T0-T7 issues a 32-bit read request for their respective values stored in registers R0, R9, R14 and R7 (thirty-two different 32-bit values in total, since four 32-bit values are read for each of eight threads). Accordingly, the crossbar 420 is configured to allow 32-bit values to be read from eight different R0 registers that each corresponds to a respective one of the threads T0-T7, which is illustrated by event 602. The crossbar 420 is also configured to allow 32-bit values to be read from eight different R9

registers that each correspond to a respective one of the threads T0-T7. The crossbar 420 is also configured to allow 32-bit values to be read from eight different R14 registers that each correspond to a respective one of the threads T0-T7. The crossbar 420 is further configured to allow 32-bit values to be read from eight different R7 registers that each correspond to a respective one of the threads T0-T7.

Thus, in one clock cycle, eight different threads T0-T7 are each capable of reading four different 32-bit registers R0, R9, R14 and R7. Subsequently, the 32-bit values extracted in event 602 are stored in operand collector 440 and eventually passed to ALU 452 for arithmetic. In some cases, however, threads T0-T7 may receive an instruction to operate on 64-bit values stored within the local register file 304. Accordingly, the manner in which the registers R0-R23 are read by the threads T0-T7 follows the techniques described below in conjunction with FIG. 7.

FIG. 7 illustrates 64-bit reads of registers R8, R10, R4 and R22 for eight threads T0-T7, according to one embodiment of the present disclosure. In the example illustrated in FIG. 7, each of the threads T0-T7 issues a 64-bit read request for their respective values stored in registers R8, R10, R4 and R22. In this case, since the read requests are 64-bit, which does not map to the 32-bit length of each of the registers R0-R23, the crossbar 420 is configured to also read the registers R9, R11, R5 and R23 for each of the threads T0-T7 along with the registers R8, R10, R4 and R22 for each of the threads T0-T7. As described in detail below, reading the registers R8-R9, R10-R11, R4-R5 and R22-R23 for each of the threads T0-T7 is performed across two clock cycles and is illustrated by the events 702 and 704.

In particular, the first read event 702 occurs in response to receiving a 64-bit read request of the registers R8, R10, R4 and R22 for the threads T0-T7, and involves reading the registers R8-R10, R9-R11, R4-R22 and R5-R23 for each of the threads T0-T3 from the banks 422, 424, 426 and 428. As shown, the 32-bit values for each of the registers R8, R10, R9, R11, R4, R22, R5 and R23 are extracted by crossbar 420 during a first clock cycle. Subsequently, for each of the threads T0-T3, the corresponding register R8 is concatenated with the register R9 to result in a 64-bit value for the register R8, the corresponding register R10 is concatenated with the register R11 to result in a 64-bit value for the register R10, the corresponding register R4 is concatenated with the register R5 to result in a 64-bit value for the register R4, and the corresponding register 22 is concatenated with register 23 to result in a 64-bit value for the register R22. This is illustrated by the arrows within the first read event 702. Thus, at the termination of the first clock cycle, the 64-bit values requested by threads T0-T3 have been extracted from the local register file 304.

Next, the second read event 704 occurs in response to receiving a 64-bit read request of the registers R8, R10, R4 and R22 for the threads T0-T7, and involves reading the registers R8-R10, R9-R11, R4-R22 and R5-R23 for each of the threads T4-T7 from the banks 422, 424, 426 and 428. As shown, the 32-bit values for each of the registers R8, R10, R9, R11, R4, R22, R5 and R23 are extracted by crossbar 420 during a second clock cycle. Subsequently, for each of the threads T4-T7, the corresponding register R8 is concatenated with the register R9 to result in a 64-bit value for the register R8, the corresponding register R10 is concatenated with the register R11 to result in a 64-bit value for the register R10, the corresponding register R4 is concatenated with the register R5 to result in a 64-bit value for the register R4, and the corresponding register 22 is concatenated with register 23 to result in a 64-bit value for the register R22.

This is illustrated by the arrows within the first read event **702**. Thus, at the termination of the second clock cycle, the 64-bit values requested by threads **T4-T7** have been extracted from the local register file **304**.

Thus, in each clock cycle, four different threads (**T0-T3/** **T4-T7**) are each capable of reading four different 64-bit registers (e.g., **R8, R10, R4** and **R22** (and the corresponding registers **R9, R11, R5** and **R23**, respectively)). Subsequently, the 64-bit values extracted in events **702** and **704** for the threads **T0-T7** are stored in operand collector **440** and eventually passed to ALU **452** for arithmetic. In some cases, however, threads **T0-T7** may receive an instruction to operate on 128-bit values stored within the local register file **304**. Accordingly, the manner in which the registers **R0-R23** are read by the threads **T0-T7** follows the techniques described below in conjunction with FIG. 7.

FIG. 8 illustrates 128-bit reads of the register **R4** for eight threads **T0-T7**, according to one embodiment of the present disclosure. In the example illustrated in FIG. 8, each of the threads **T0-T7** issues a 128-bit read request for their respective values stored in register **R4**. In this case, since the read requests are 128-bit, which does not map to the 32-bit length of the registers **R0**, the crossbar **420** is configured to allow the registers **R5, R6** and **R7** to be read by each of the threads **T0-T7** in addition to allowing the register **R4** to be read by each of the threads **T0-T7**. The 32-bit values of registers **R4, R5, R6** and **R7** are then concatenated to establish the respective 128-bit value that is requested by each of the threads **T0-T7**, as described in further detail below.

Event **802** illustrates 128-bit reads of register **R4** for threads **T0-T3** during a first clock cycle. As mentioned above, when a 128-bit read request of a register is requested, the crossbar **420** is configured to automatically read the three registers that immediately follow the register, i.e., registers **R5-R7** for a 128-bit read request of register **R4**. As illustrated, in event **802**, the 32-bit values of registers **R4-R7** are reordered such that the 128-bit values are properly ordered (**R4-R7**) for each of the threads **T0-T3**.

Event **804** illustrates 128-bit reads of register **R4** for threads **T4-T7** during a second clock cycle. According to the techniques described above, the crossbar **420** is configured to automatically read the three registers **R5-R7** in addition to the register **R4** for the threads **T4-T7** in order to extract the respective 128-bit values. Again, the 32-bit values of registers **R4-R7** are reordered such that the 128-bit values are properly ordered (**R4-R7**) for each of the threads **T4-T7**.

Thus, in each clock cycle, four different threads (e.g., **T0-T3**) are each capable of reading a corresponding 128-bit register (e.g., **R4-R7**). Subsequently, the 128-bit values extracted in events **802** and **804** for the threads **T0-T7** are stored in operand collector **440** and eventually passed to ALU **452** for arithmetic.

One embodiment of the disclosure may be implemented as a program product for use with a computer system. The program(s) of the program product define functions of the embodiments (including the methods described herein) and can be contained on a variety of computer-readable storage media. Illustrative computer-readable storage media include, but are not limited to: (i) non-writable storage media (e.g., read-only memory devices within a computer such as compact disc read only memory (CD-ROM) disks readable by a CD-ROM drive, flash memory, read only memory (ROM) chips or any type of solid-state non-volatile semiconductor memory) on which information is permanently stored; and (ii) writable storage media (e.g., floppy disks within a diskette drive or hard-disk drive or any type

of solid-state random-access semiconductor memory) on which alterable information is stored.

The disclosure has been described above with reference to specific embodiments. Persons of ordinary skill in the art, however, will understand that various modifications and changes may be made thereto without departing from the broader spirit and scope of the disclosure as set forth in the appended claims. The foregoing description and drawings have, accordingly, to be regarded in an illustrative rather than a restrictive sense.

What is claimed is:

1. A computer-implemented method for performing access of a register file that includes a set of N registers, wherein N is greater than or equal to two, the method comprising:

from each thread included in a plurality of threads executed by a processor, receiving a request to access a first amount of data from each register in the set of N registers; and

configuring a crossbar, connecting the register file with the processor, to allow the plurality of threads to access data from each register in the set of N registers, and to allow each thread of the plurality of threads to access the first amount of data from each register in the set of N registers in a single clock cycle.

2. The computer-implemented method of claim 1, wherein the register file comprises a plurality of memory banks, and each register in the set of N registers resides in a different memory bank.

3. The computer-implemented method of claim 1, wherein, at each clock cycle, each thread accesses the first amount of data from either:

each register in the set of N registers to produce an access operation comprising N accesses of the first amount of data;

each register in $N/2$ aligned pairs of sequential registers to produce an access operation comprising $N/2$ accesses of twice the first amount of data; or

each register in $N/4$ aligned quads of sequential registers to produce an access operation comprising $N/4$ accesses of four times the first amount of data.

4. The computer-implemented method of claim 1, wherein the set of N registers includes $N/2$ aligned pairs of sequential registers, and each thread accesses the first amount of data from each register in each aligned pair of sequential registers in a single clock cycle to produce a shaped access operation comprising $N/2$ accesses of twice the first amount of data.

5. The computer-implemented method of claim 1, wherein the set of N registers includes $N/4$ aligned quads of sequential registers, and each thread accesses the first amount of data from each register in each aligned quad of sequential registers in a single clock cycle to produce a shaped access operation comprising $N/4$ accesses of four times the first amount of data.

6. The computer-implemented method of claim 1, wherein the plurality of threads comprises a subset of a second group of threads and includes either a sequential lower half of the threads included in the second group of threads or a sequential upper half of the threads included in the second group of threads.

7. The computer-implemented method of claim 1, wherein each thread accesses the first amount of data by either reading the first amount of data from each register in the set of N registers or writing the first amount of data to each register in the set of N registers.

17

8. The computer-implemented method of claim 1, wherein the first amount of data comprises 32-bits of data.

9. A non-transitory computer-readable storage medium storing instructions that, when executed by a processor, cause the processor to perform access of a register file that includes a set of N registers, wherein N is greater than or equal to two, by performing the steps of:

from each thread included in a plurality of threads executed by a processor, receiving a request to access a first amount of data from each register in the set of N registers; and

configuring a crossbar, connecting the register file with the processor, to allow the plurality of threads to access data from each register in the set of N registers, and to allow each thread of the plurality of threads to access the first amount of data from each register in the set of N registers in a single clock cycle.

10. The non-transitory computer-readable storage medium of claim 9, wherein the register file comprises a plurality of memory banks, and each register in the set of N registers resides in a different memory bank.

11. The non-transitory computer-readable storage medium of claim 9, wherein, at each clock cycle, each thread accesses the first amount of data from either:

each register in the set of N registers to produce an access operation comprising N accesses of the first amount of data;

each register in N/2 aligned pairs of sequential registers to produce an access operation comprising N/2 accesses of twice the first amount of data; or

each register in N/4 aligned quads of sequential registers to produce an access operation comprising N/4 accesses of four times the first amount of data.

12. The non-transitory computer-readable storage medium of claim 9, wherein the set of N registers includes N/2 aligned pairs of sequential registers, and each thread accesses the first amount of data from each register in each aligned pair of sequential registers in a single clock cycle to produce a shaped access operation comprising N/2 accesses of twice the first amount of data.

13. The non-transitory computer-readable storage medium of claim 9, wherein the set of N registers includes N/4 aligned quads of sequential registers, and each thread accesses the first amount of data from each register in each aligned quad of sequential registers in a single clock cycle to produce a shaped access operation comprising N/4 accesses of four times the first amount of data.

14. The non-transitory computer-readable storage medium of claim 9, wherein the plurality of threads comprises a subset of a second group of threads and includes

18

either a sequential lower half of the threads included in the second group of threads or a sequential upper half of the threads included in the second group of threads.

15. The non-transitory computer-readable storage medium of claim 9, wherein each thread accesses the first amount of data by either reading the first amount of data from each register in the set of N registers or writing the first amount of data to each register in the set of N registers.

16. The non-transitory computer-readable storage medium of claim 9, wherein the first amount of data comprises 32-bits of data.

17. A system for performing access of a register file that includes a set of N registers, wherein N is greater than or equal to two, the system comprising:

the register file;

a crossbar connecting the register file with a processor; and

the processor configured to:

from each thread included in a plurality of threads executed by a processor, receive a request to access a first amount of data from each register in the set of N registers; and

configure the crossbar to allow the plurality of threads to access data from each register in the set of N registers, and to allow each thread of the plurality of threads to access the first amount of data from each register in the set of N registers in a single clock cycle.

18. The system of claim 17, wherein the register file comprises a plurality of memory banks, and each register in the set of N registers resides in a different memory bank.

19. The system of claim 17, wherein, at each clock cycle, each thread accesses the first amount of data from either:

each register in the set of N registers to produce an access operation comprising N accesses of the first amount of data;

each register in N/2 aligned pairs of sequential registers to produce an access operation comprising N/2 accesses of twice the first amount of data; or

each register in N/4 aligned quads of sequential registers to produce an access operation comprising N/4 accesses of four times the first amount of data.

20. The system of claim 17, wherein the set of N registers includes N/2 aligned pairs of sequential registers, and each thread accesses the first amount of data from each register in each aligned pair of sequential registers in a single clock cycle to produce a shaped access operation comprising N/2 accesses of twice the first amount of data.

* * * * *